

Hierarchical Clustering for thematic browsing and summarization of large sets of Association Rules *

Alípio Jorge[†]

Abstract

In this paper we propose a method for grouping and summarizing large sets of association rules according to the items contained in each rule. We use hierarchical clustering to partition the initial rule set into thematically coherent subsets. This enables the summarization of the rule set by adequately choosing a representative rule for each subset, and helps in the interactive exploration of the rule model by the user. We define the requirements of our approach, and formally show the adequacy of the chosen approach to our aims. Rule clusters can also be used to infer novel interest measures for the rules. Such measures are based on the lexicon of the rules and are complementary to measures based on statistical properties, such as confidence, lift and conviction. We show examples of the application of the proposed techniques.

1 Introduction.

Despite being popular as a technique for market basket analysis, association rules [1][26] are now used in many different applications, from modeling web user preferences [9], to studying census data [6]. The apriori algorithm [2], and variants [6][20][23], among others, are the standard technique for association rule discovery. The mining process, however, is not finished when the rules are produced. A set of association rules is mostly a descriptive model that typically requires post processing before actionable information (information that can be acted upon in order to produce value [5]) is found. Moreover, due to the completeness of the rule discovery algorithm, the set of rules generated for a single problem can be very large, easily reaching hundreds or even thousands of rules [13].

Post processing techniques mainly encompass rule filtering (or pruning), using statistical measures of interest [6][13][22], rule set querying using SQL like lan-

guages [7][8][17], rule summarization, by finding a relevant subset of the rules [4][13], and rule browsing in interactive environments, that provide a more intuitive interface and let the user combine the objectiveness of statistical measures with subjective domain knowledge [10][11][16]. The aim, in any case, is to allow the user the inspection of smaller relevant subsets of the whole set of rules. The questions are: how to choose the subsets, how to decide what is relevant and what is not, how to structure the rule space so that the user/data analyst can have a digestible view of the whole rule set, and explore the subsets that seem of interest.

In this paper we explore the lexical features of the rules, rather than their statistical properties, for structuring the rule space, dividing the rule model into coherent thematic subsets and producing small summary rule sets. We perform hierarchical clustering to partition the set of rules at various levels, to select a set of representative rules and to define a new interest measure that expresses the rules lexical singularity.

We start by giving a more detailed account of the possibilities in association rule post processing. We then explain the intuition behind our approach for rule clustering and summarization, and after that we define our formal framework. Hierarchical clustering is briefly summarized and then applied to the summarization of association rules. We formally justify our choice of the hierarchical clustering algorithm and show some examples of its application. After some further discussion we relate our approach to rule browsing environments and to other related work, and conclude.

2 Post processing of Association Rules

The general aim of data mining is to find patterns in data that can become actionable information [5]. In the case of association discovery, the involvement of domain experts is critical to the process of identification of relevant rules. The exploration of a rule set is ideally an interactive process, providing objective measures of the interest of each rule, but also giving room to less tangible interests related to the domain of the data set. We call this non-trivial analysis process, that starts at the rule model and ends up with a relevant subset of

*Supported by the POSI/SRI/39630/2001/Class Project (Fundação Ciência e Tecnologia), PortalExecutivo S.A., www.portalexecutivo.com, POE/SIME/80/00666 (Ministério da Economia), FEDER e Programa de Financiamento Plurianual de Unidades de I & D.

[†]LIACC, Faculdade de Economia do Porto, R. do Campo Alegre 823, 4150-180 Porto, Universidade do Porto, Portugal, amjorge@liacc.up.pt, <http://www.niaad.liacc.up.pt/~amjorge>.

rules, the *post processing of association rules*.

Previous and current work on association rule post processing can be divided into four main streams.

Finding rules with outstanding statistical properties: this is achieved by characterizing each rule with some chosen measures and looking for rules above or below given thresholds. The most popular measures are support, confidence, interest (or lift) [1], leverage, conviction [6] and χ^2 [13]. Most of these measures are concerned with the detection of rules $a \rightarrow b$, where a and b deviate from statistical independence. A more generalized view of this approach [21] regards the above as objective interestingness measures, and refers to subjective measures of interest as the ones that characterize each rule on how the users personal expectations are met.

Rule set querying: the analyst can explore the rule set through the use of a query language (typically inspired in SQL) [7][8][17]. Using such a querying facility it is then possible to obtain subsets of rules containing certain items, and/or satisfying some user provided constraints involving support and confidence, for example.

Finding representative or optimal subsets of rules (summarizing): optimal subsets of rules can be found for some of the interest measures referred above [4]. The rules that are not optimal can be filtered out. This technique does not require the user to provide thresholds. Another approach to summarization is the identification of *direction setting* rules [13]. A direction setting rule $a \rightarrow b$ is a relatively small rule that represents larger rules $a' \rightarrow b$ ($a \subseteq a'$) with the same *direction*, where direction is dictated by the status of statistical independence of antecedent and consequent, as measured by the χ^2 test. A large rule set can be summarized by a relatively small set of direction setting rules.

Browsing and visualizing rules: association rule browsing environments [10][11][16], help the user in the exploration of a large set of association rules either by dividing the initial set of rules into smaller digestible sets, starting from a summary set of rules, or by providing graphical representations of the rule set or subsets. The user can interactively navigate through the rule set, inspecting subsets of rules, obtaining the values of various measures for each rule, and so on.

Our present work presents contributions to the last two items (rule summarization and rule browsing). We propose a technique for clustering association rules according to items occurring in the rules. This will produce thematic groups that can be analyzed independently of each other. Given a set of clusters,

we can summarize the whole rule set by finding an appropriate representative rule for each cluster. Such a rule can be used as a starting point for exploring the corresponding cluster.

3 Item-based clustering of rules

The central idea of our approach is the application of hierarchical clustering to sets of association rules. The rules are clustered according to the items they contain, rather than any relationship with the data. This gives the data analyst groups of rules corresponding to different themes in the domain. Such a thematic analysis can be performed manually without much effort if the rule set is sufficiently small and if each rule clearly corresponds to a different theme. Typically, this is not the case: rule sets can be very large, and the thematic separation of rules may not be clear. As a simple example, in Table 1 we can see a very small rule set with items referring to taxes, government, books on management and international politics. The rules represent associations between keywords in news articles for business executives. Some rule subsets are clearly separable ($\{r_{11}, r_{12}, r_{13}\}$ and $\{r_{14}, r_{15}, r_{16}\}$) because they belong to different themes with no common items. The first 10 rules, however, can be one single group, or can be divided into more than one group depending on the criteria used. We will also see, in this paper, what these criteria are.

Table 1: *Keywords small rule set*

r1 :	customs & imports -> taxes
r2 :	customs -> taxes
r3 :	imports -> taxes
r4 :	customs & imports & tax payer -> taxes
r5 :	customs & tax payer -> taxes
r6 :	taxes & imports -> customs
r7 :	taxes -> customs
r8 :	crisis -> taxes
r9 :	crisis & government -> taxes
r10:	taxes -> crisis
r11:	management -> books
r12:	books -> management
r13:	management & success -> books
r14:	Iraq & USA -> Europe
r15:	Iraq -> USA
r16:	USA -> Iraq

After dividing the rule set into a number of subsets, we can explore the association rule model in a more systematic way. Each rule subset can be analyzed separately. If the rule subset is still too large, we can further divide it into subsets. If the rules in one subset

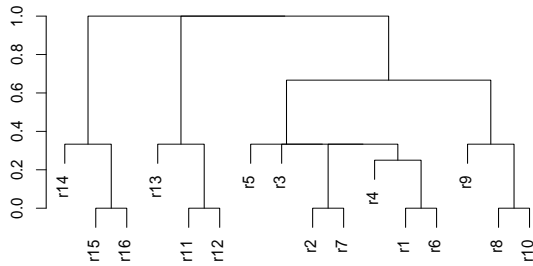


Figure 1: Dendrogram for the keywords small set of rules. The scale at the left shows the height of the aggregation.

share many items, we can choose one of the rules from that subset to serve as a representative of that group. The set of the representative rules of the various subsets can be presented to the data analyst as a summary of the whole rule set.

In Figure 1 we can see the dendrogram representing the hierarchical clustering, obtained with the statistical tool R1.7 [12][18], for the *keywords small* rule set shown in Table 1. This clustering tree shows the aggregation of rules at different similarity levels (*heights*). More similar rules are joined at lower heights. The height corresponds to the distance between the two joined clusters.

We can obtain a given number k of clusters by cutting the clustering tree appropriately. For $k = 3$ and $k = 4$ we have a good thematic separation of the rules in the example. For other values of k we get some arbitrary and apparently meaningless separations. Rules {r11, r12, r13} are about books on management, rules {r14, r15, r16} are about international politics. The remaining rules can be one single cluster on taxes ($k = 3$), or can have a sub-cluster {r8, r9, r10} on government ($k = 4$).

This simple example illustrates some of the issues we are dealing with. We can divide a set of rules into subsets that can be analyzed independently of each other; thematic subsets of rules may or may not be clearly separable; the choice of the number of clusters is important. Some fundamental questions are: what do we mean by *clearly separable* and by *thematic sets of rules*? In the remaining of the paper we will formalize these notions, provide more details about the clustering process and options, give criteria for the choice of the number of clusters k , and look into some applications of thematic rule clustering.

4 Framework

A good thematic cluster contains rules that share items with each other but not with rules in other clusters. In a real rule set, it may not be convenient or even possible to obtain a set of clusters with such properties. In fact, we may be interested in a specific number k of clusters, say 20, so that we can present to the user a summary of the rule set that fits into a page or an application window. For an arbitrary k , different clusters will probably share items. The opposite may also occur: some clusters may be partitionable into subsets that have no items in common. This is clearly undesirable.

To reason about the rule set, we can view it as a graph (the *thematic graph*). Each rule is a node, with direct links to nodes that share at least one item. In such a setting, a *separable thematic* subset of rules corresponds to a subgraph that has no links to the rest of the graph. An *integer thematic* set of rules is a set that does not strictly contain any separable thematic subset. Two rules r and s are *thematically linkable* if there is a path in the thematic graph from r to s . Two clusters A and B are *thematically linkable* if at least one rule in A is thematically linkable to another in B .

DEFINITION 4.1. *Given a set of rules R , a clustering $C = \{C_1, C_2, \dots, C_k\}$ of R is thematically canonical if:*

- a) *All C_i are separable thematic subsets.*
- b) *All C_i are integer thematic subsets.*

Definition 4.1 provides very precise requirements for an ideal rule set clustering, where a clustering of a rule set R is a set of disjoint and complementary subsets of R . Under such conditions, no two clusters C_i, C_j are thematically linkable. At the same time, for every cluster C_i , every rule in that cluster is thematically linkable to any other rule in the same cluster.

Any non-empty given set of rules can be partitioned into exactly $K \geq 1$ thematic subsets that are both separable and integer. The precise value of K can be algorithmically obtained. If the rule set is divided into an arbitrary number k of clusters and if k is smaller than K , then it is possible to satisfy condition a) only. If $k > K$ then condition b) can be satisfied. Definition 4.2 establishes the conditions that any clustering of arbitrary size k must satisfy with respect to the canonical clustering of size K .

DEFINITION 4.2. *Given a set of rules R , let K be the size of its canonical clustering. A clustering $C = \{C_1, C_2, \dots, C_k\}$ is thematically admissible if:*

- *for $k = K$, C is thematically canonical;*
- *for $k < K$, C satisfies condition a) of Definition 4.1;*
- *for $k > K$, C satisfies condition b) of Definition 4.1.*

Now we have the minimal requirements for the clustering of a rule set. This will guide us in the search of an appropriate clustering method. As we will see in §7, thematic admissibility can be ensured if we choose appropriate rule distance metric and clustering strategy.

The adopted distance between two rules r and s will be the *binary* distance metric.

$$(4.1) \text{dist}(r, s) = 1 - \frac{\{\text{items in } r\} \cap \{\text{items in } s\}}{\{\text{items in } r\} \cup \{\text{items in } s\}}$$

The binary distance between two rules is proportional to the number of items that occur in only one of the rules. The rule graph can be enriched by labeling the edges with the respective distance value between the two nodes. In this case, a distance based clustering algorithm will tend to minimize the average number of items shared by different clusters and maximize the number of items shared within each cluster.

5 Hierarchical clustering

After obtaining the rule set R , we build a distance matrix D that contains the distance between rules r_i and r_j in entry $D_{i,j}$. We use the binary distance as defined in (4.1). Other distance metrics are discussed in §10. The matrix D is then used to build the hierarchical clustering tree T . This tree can be cut into a given number of clusters. Different strategies to build the tree result in different clusterings for the same number of clusters. All these steps have been performed using the statistical platform R1.7 [18].

We use a standard bottom-up hierarchical clustering algorithm to build the clustering tree (Algorithm 5.1)[8]. It starts with each rule as a singleton cluster, and proceeds bottom up by iteratively joining the two nearest clusters.

ALGORITHM 5.1. *Bottom-up hierarchical clustering*

1. Let each rule in R be a singleton cluster;
2. If the number of clusters is 1 then stop;
3. Join the two nearest clusters;
4. Go to step 2.

To decide which are the two nearest clusters to join we need to define a distance metric for clusters as well. This will be stated in terms of the distance between the rules belonging to different clusters. The choice of the cluster distance metric determines the clustering strategy and therefore the properties of the resulting clusters.

The distance between two clusters C_1 and C_2 we have adopted is $\text{dist}C_{avg}(C_1, C_2)$, corresponding to the

average distance between two rules of different clusters (see justification in §7).

$$\text{dist}C_{avg}(C_1, C_2) = \frac{1}{\#C_1 \cdot \#C_2} \sum_{r \in C_1} \sum_{s \in C_2} \text{dist}(r, s)$$

We will also discuss and compare other distance metrics between clusters. The minimal distance, which corresponds to using the strategy *single linkage*, also called *nearest-neighbor*, for aggregating the clusters is defined as:

$$\text{dist}C_{sgl}(C_1, C_2) = \min_{r \in C_1, s \in C_2} \text{dist}(r, s)$$

The maximal distance (*complete linkage* strategy) is obtained by using the *max* function instead of the *min* function.

$$\text{dist}C_{cpl}(C_1, C_2) = \max_{r \in C_1, s \in C_2} \text{dist}(r, s)$$

For any of the three strategies, when two clusters are aggregated, their inter-cluster distance corresponds to the *height* at which their aggregation appears on the dendrogram (tree representing the hierarchical clustering).

The maximal and the average strategies tend to form compact clusters, i.e., clusters where all the elements in the same cluster are relatively close to each other. Clusters formed with the minimal strategy tend to be less compact. In particular, a cluster can represent a chain of nearest neighbors, where the first element of the chain is very distant from the last one. In §7 we will study in more detail the properties of the clustering strategies that are more adequate to post processing association rules.

6 Summarizing sets of rules

One explored technique for handling large sets of rules is to find a subset of representative rules that can give the user a rough description of the data. Such a subset of representatives can also be used as a starting point for the exploration of the whole rules set. We will call such a subset a *rule set summary* and will identify its important features.

Going back to the *keywords small* example, a rule set summary, for 4 clusters, is shown in Table 2.

After obtaining a rule clustering, for a given number of clusters k , we choose a representative rule for each cluster. By looking at the rule, the analyst can have an idea about the theme of the cluster it represents. With appropriate tools (such as PEAR [10], or SQL-like data mining languages, [7], [8], [17]), the analyst can *mine around a chosen rule*, i.e., fetch thematically similar

Table 2: Summary of *keywords small* rule set

representative rule	cluster size
r1 : customs & imports -> taxes	7
r8 : crisis -> taxes	3
r11: management -> books	3
r15: Iraq -> USA	3

rules that have some items in common. We can make sure that the representative of a rule set is thematically linkable to all the rules in the set. This is guaranteed if the rule set is thematically integer, which is in turn guaranteed in our approach if we choose a number of clusters $k \geq K$, where K is the number of separable integer subsets of the rule set, as we will prove in §7.

To be a good starting point for further rule exploration, the representative of an integer rule set should also minimize the average effort of the analyst for exploring the rule set. This effort depends on the post processing tools available. It seems reasonable, however, that a rule set exploration tool relates more easily rules with more items in common. Therefore, we choose the representative $rep(R)$ of a rule set R as the rule r that has the minimal average binary distance to all the other rules in the set. If more than one rule minimizes the average distance, one of them is arbitrarily chosen.

$$rep(R) = \arg \min_{r \in R} [average_{s \neq r \in R} dist(r, s)]$$

To summarize a rule set, we choose the number k of representative rules we want. This is typically dictated by the size of the page or of the window we will use for display. We produce k clusters, so that each cluster is integer. The summary is the set of representative rules of each of the k clusters.

7 Theoretical justification

In this section we study some of the theoretical aspects of thematic hierarchical clustering. In particular, we justify the choice of the average strategy, in combination with the binary distance between rules, for cluster aggregation. Another important issue is the determination of an appropriate number of clusters. We show how this number can be found.

THEOREM 7.1. *Hierarchical clustering of association rules, using the binary distance and the single linkage strategy is thematically admissible.*

Proof. Let K be the number of separable and integer subsets of a rule set R . The single linkage bottom up strategy only joins two separable clusters if all other clusters are separable. Two separable clusters have a binary distance of 1. Two non-separable clusters have a

binary distance of $d < 1$. Therefore, the last K clusters left to join are the K separable ones. These clusters, and only these ones are aggregated at height 1.

Cutting the tree to obtain $k = K$ clusters results precisely in the K separable clusters. Each of these clusters is also integer, otherwise we would have more than K separable clusters.

Cutting the tree to obtain $k < K$ clusters necessarily results in separable clusters, since the aggregation of two separable clusters is a separable cluster.

Cutting the tree to obtain $k > K$ clusters necessarily results in integer clusters since the aggregation of the subsets of the K clusters was done at a height lower than 1.

THEOREM 7.2. *Hierarchical clustering of association rules, using the binary distance and the average linkage strategy is thematically admissible.*

Proof. The average distance between two clusters is 1 if and only if the minimal (single) distance is 1. Therefore, Theorem 7.2 is proved similarly to Theorem 7.1.

As described in the proof of Theorem 7.1, we can easily obtain the $k = K$ clusters, by cutting the clustering tree just below height 1 (and above the second lowest observed height). This applies to both single and average linkage. As a consequence, given a set of rules R , we can always identify the number K corresponding to the size of the thematically canonical clustering of those rules.

There is a simple graph traversing algorithm, $O(n^2)$ that identifies the K clusters. The advantage of using hierarchical clustering is that, after determining the value for K , we can choose the number of clusters k which is more convenient for rule summarization and presentation. As we have shown, when $k = K$, we obtain exactly the integer and separable clusters, as a simple graph traversing algorithm would. If $k < K$, we get separable clusters, i.e., clusters that have no items in common. Necessarily, at least one of these clusters will be non-integer. In this case, it is not possible to pick one representative rule for the whole cluster. When $k > K$, we get integer clusters, which means that we can pick any rule as the representative of the respective cluster.

It is possible to find counter examples that show that the *complete linkage + binary distance* combination is not thematically canonical. For a number $k = K$ of clusters, we may obtain non-separable and non-integer clusters. Nevertheless, it is true that for the complete linkage, if we cut the clustering tree just below height 1 we always obtain integer clusters.

For integer sets of rules ($k > K$), *complete linkage* and *average linkage* give more compact clusters than

single linkage. In a compact cluster, the number of common rule items is higher.

The adequate choice for the clustering method is thus the *average linkage*. It identifies the separable integer subsets of rules, if any, and it gives compact clusters when no separable subsets exist, which makes it suitable for iterative application.

8 Empirical study

In this Section we show two case studies that illustrate our approach. The rules have been obtained with Caren [3], a java implementation of APRIORI, with extensions. For post processing, we have used R1.7 (mainly functions `dist`, `hclust`, `cutree`, and own-built functions). Clustering was done with the *average* strategy and the *binary* distance.

8.1 Keywords (94 rules) This example is taken from a study of searchable keywords of published executive news texts of a web portal¹. Keywords are chosen by the news editors at the time of publication. Each text has 1 to 4 keywords. The ideal K is 7. This can be easily determined after building the clustering tree. The clusters found at height just below 1 (and above 0.667, the next observed distance between clusters in this case), as well as the representative rules and cluster sizes are shown in Table 3. We also suggest a name for each cluster/theme.

These clusters share no items with each other (they are separable). Moreover, each one of the clusters has no separable subsets (they are integer). If we cut the clustering tree at a lower height (between the two observed distances 0.5 and 0.667), we obtain 29 clusters.

Table 3: Canonical thematic clusters for a real keywords rule set.

	representative	size	theme
1	Taxes on individual income & Taxes on corporate income -> Fiscality	27	taxes
2	New members & Europe -> European Union	23	politics
3	Business administration & Management -> Training	24	training
4	Results -> Financial statements	2	finance
5	Books -> Strategy	14	books
6	Links -> Markets	2	markets
7	Gross National Product -> Recession	2	recession

¹<http://www.portalexecutivo.com>

8.2 Reuters keywords (501 rules) This is also a data set with keywords of news articles but from a different source: the Reuters news agency [15]. Our approach determined that the number of integer and separable clusters is 4 (Table 4).

Table 4: Summary rule set for the Reuters keywords rule set: the 4 thematically canonical clusters.

	representative	size
1	grain & wheat & soybean & oilseed -> corn	495
2	gold -> silver	2
3	coffee -> cocoa	2
4	zinc -> lead	2

We can observe a very unbalanced size distribution of the clusters, nevertheless reflecting quite reasonable thematic divisions (first need goods, precious metals, imported food commodities and other metals). Note that this unbalance is a feature of the rule set, rather than of the approach. Nevertheless, such a summary rule set is of little use.

For 20 clusters (arbitrary value), the sizes are more balanced (Table 5).

Table 5: Summary rule set for the Reuters keywords (20 clusters).

	representative	size
1	grain & wheat & oilseed -> cotton	74
2	money-fx -> dlr	28
3	gold -> silver	2
4	grain & sorghum & soybean & oilseed -> corn	149
5	grain & soy-meal & meal-feed -> corn	61
6	wheat & oat -> grain	12
7	veg-oil & oilseed & soy-oil -> soybean	74
8	rice -> grain	25
9	livestock & grain -> carcass	9
10	coffee -> cocoa	2
11	grain & oilseed -> sunseed	17
12	dlr & money-fx -> interest	12
13	ship -> grain	5
14	-> trade	7
15	-> acq	2
16	ship -> crude	10
17	fuel -> gas	2
18	gnp -> bop	6
19	-> earn	2
20	zinc -> lead	2

The large cluster of Table 4 has been split into 17 new sub-clusters. We now can see some of the

corresponding sub-themes: currency (clusters 2 and 12 of Table 5); vegetable oils (cluster 7); livestock (9); crude shipping (16) and so on.

Cluster number 4 of Table 5, which is still quite large, can subsequently be divided into 20 more clusters (Table 6). Thematic distinctions are now hard to see for the untrained eye.

Table 6: Sub-clustering of cluster 4 in Table 5.

	representative	size
1	grain & sorghum & barley -> corn	12
2	grain & sorghum & soybean & oilseed -> corn	16
3	grain & sorghum & oilseed -> corn	12
4	grain & wheat & sorghum -> corn	12
5	grain & wheat & soybean & oilseed -> corn	16
6	wheat & soybean & oilseed -> corn	12
7	grain & wheat & oilseed -> corn	7
8	grain & wheat & barley -> corn	7
9	grain & wheat -> corn	5
10	wheat & sorghum -> grain	5
11	wheat & soybean & oilseed -> grain	7
12	sorghum & barley -> grain	5
13	sorghum & soybean & oilseed -> grain	7
14	wheat & barley -> grain	5
15	soybean & oilseed -> grain	5
16	barley -> sorghum	2
17	soybean & oilseed -> sorghum	5
18	wheat -> sorghum	2
19	wheat & oilseed -> soybean	5
20	wheat -> soybean	2

We should bear in mind that the rules in a summary set are merely starting points for the exploration of the rule set. We can start from one of the representative rules and explore the corresponding sub-cluster, but the post-processing environment should also allow browsing through similar rules in other clusters.

A rule set may have no separable subsets. This happens more easily if the rules are produced with very low support and confidence (rare combinations appear and link two or more separable rule sets). A rule set with more than 2500 rules, obtained with the Reuters data used above had no separable subsets. In that case, any k clusters will be separable. Hierarchical thematic clustering still provides useful results in such conditions, since the distance metric will allow the identification of meaningful clusters.

But what happens if we have a canonical clustering of size K for a given set of rules R , and we then obtain a superset $R' \supset R$ using a lower minimal support? Which clusters would we obtain from R' using the same value of K ? Would the cluster themes change radically? Would

the original clusters still be found as subsets in the new clustering? Not necessarily. This suggests a possible improvement of our method by taking the support of the rules into account, so that a change in the minimal support of the rules produces, as much as possible, monotonic changes in their thematic clustering. This possibility is currently being investigated.

Another important issue is related to clusters with overlapping items. A given rules may have items in common with more than one cluster. It could make sense to have the same rule belonging to different clusters. However, this is not a serious drawback for our current aims.

9 Measuring the distinctness of rules

We can use the above described notions of distance and theme to characterize each rule of a set R in terms of its thematic distinctness. This can be used as a measure of interest that is not related to the usual statistical measures such as support, confidence, lift or conviction. The thematic distinctness corresponds to the degree to which each rule is lexically different from the others.

Rules with high thematic distinctness are rules with rare associations, although the inverse is not necessarily true. The distinctness measure can be used for example to identify rules that are not variations of other existing rules.

A straightforward way of measuring the thematic distinctness of a rule is the number of rules in the cluster. Small clusters will correspond to less frequent combination of items and therefore to possibly interesting rules. Such a measure, however, would depend on the number of clusters chosen. We will, instead, use the *binary* distance for defining rule distinctness.

The distinctness measure will have the following properties:

- $0 < distinctness(r) \leq 1$
- $distinctness(r) = 1$ iff no rule is similar to r .
- $distinctness(r_1) < distinctness(r_2)$ iff the average similarity of other rules to r_1 is strictly larger than the average similarity of rules to r_2 .

The *distinctness* of a rule r in a rule set R can be measured as one minus the average similarity to the other rules:

$$distinctness(r) = 1 - \frac{1}{\#R - 1} \sum_{s \neq r \in R} (1 - dist(r, s))$$

This can be easily computed from the distance matrix that must to be produced for hierarchical clustering anyway. In Table 7 we can see the values of distinctness obtained for the rules of the *small keywords* set. Rules in smaller thematic clusters have higher distinctness.

Table 7: Distinctness of rules from the *keywords small* example. Rules are decreasingly sorted by their distinctness.

rule	distinct.
Iraq -> USA	0.933
USA -> Iraq	0.933
management & success -> books	0.911
Iraq & USA -> Europe	0.911
management -> books	0.889
books -> management	0.889
crisis & government -> taxes	0.810
crisis -> taxes	0.759
taxes -> crisis	0.759
imports -> taxes	0.756
customs & tax payer -> taxes	0.731
customs & imports & tax payer -> taxes	0.712
customs -> taxes	0.683
taxes -> customs	0.683
customs & imports -> taxes	0.670
taxes & imports -> customs	0.670

With very large sets of rules we may benefit in considering the k nearest neighbours of the rule.

10 Options in hierarchical clustering

To apply hierarchical clustering, we have to choose one distance metric between rules and one aggregation strategy (one distance metric between clusters). The distance metrics immediately available in R1.7 are common ones: *Euclidean*, *maximum*, *Canberra*, *Manhattan* and *binary*. As we have shown, the binary distance has good properties, and was the one chosen. The other metrics are more appropriate for numeric vectors and are not guaranteed to satisfy the requirements of thematic clustering. For the case of the Euclidean distance it is possible to find examples of poor thematic clustering.

The distance between two clusters can be measured as the maximum distance between two rules of different clusters (*single linkage*), the minimum distance (*complete linkage*) and the average distance (*average*). Once more, we chose the one that satisfied the requirements; the average distance. Some other metrics exist and may be considered in the future in comparison to the adopted one.

11 Clustering and rule browsing

Rule clustering can be used iteratively and interactively in a rule post processing environment such as PEAR (Post-processing Environment of Association Rules)[10]. PEAR is a web-based environment that assists the analyst in the exploration of the rule set, by dividing it into tangible and coherent subsets. Each

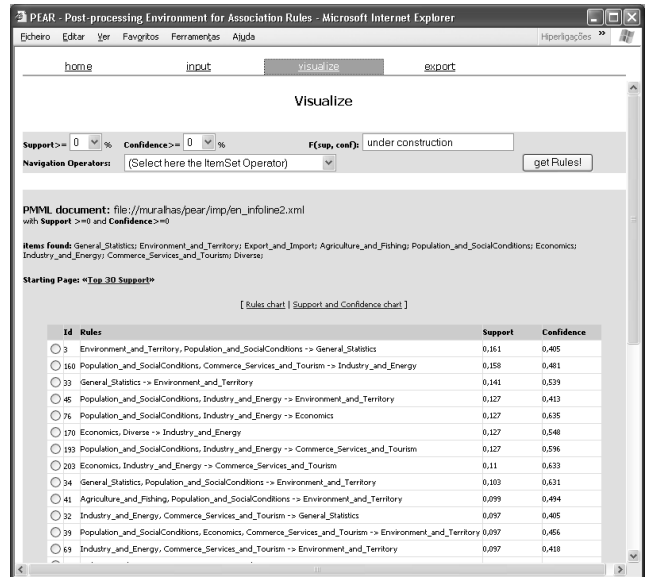


Figure 2: Main page of PEAR, a web-based environment for post-processing of association rules.

subset of rules is presented as a web page (see Figure 2), and should be small enough to fit into one screen. Besides being able to visualize some of the features of the viewed rules, the analyst can move to other subsets by applying one of a list of predefined operators to one of the viewed rules.

The exploration of the rule set starts with an initial set of rules, which serves as an index page. From here, the data analyst has an overview of the whole rule set, and can choose one of the rules as hyperlink to follow. The index page can obviously correspond to the summary set of rules provided by the clustering technique described in this paper.

When one of the rules is chosen, the analyst can apply one of a set of pre-defined operators to the initial rule, obtaining a new subset of rules. An example of such an operator is *Antecedent Generalization (AG)*. The application of the operator AG to a rule $a \rightarrow b$ results in the set of rules $\{a' \rightarrow b, \text{ such that, } a' \subset a\}$. If the resulting set is still large enough, clustering can be iteratively applied.

A high level description of the post-processing steps is shown in Algorithm 11.1.

ALGORITHM 11.1. Post-processing steps

1. produce rule set R
2. divide rule set into k clusters
3. find a representative rule for each cluster R_k
4. list the representatives of each R_k
5. chose one of the rules and explore (e.g., apply operator) the corresponding rule subset

6. if the subset is large enough (contains more than k rules) repeat from second step. Otherwise simply show the rules.

12 Comparison with related work

Thematic clustering for association rule summarization is primarily related with the already mentioned direction setting rules technique. Our clustering technique has, however, a different perspective on the rules, by exploring lexical similarities, instead of statistical features of the rules. The summarization provided by the two techniques give different views of the rule set. These views are not totally independent, since a direction setting rule also shares items with the rules it represents (while themselves have some similar statistical features). However, in thematic clustering the number of rules to be presented as a summary can, in general, be chosen. In particular, thematic clustering may even be used to summarize a large set of direction setting rules (e.g., 500 rules).

In summary, thematic clustering seems more appropriate for dividing the rule set into connected digestible pages, assuming that any rule in the rule set may be interesting (even if for subjective reasons). The DS technique exploits one particular statistical aspect of the rule set (assuming the χ^2 test is the only guiding interest measure), which, although very relevant, is necessarily restrictive.

Hierarchical clustering has been used before for grouping large sets of IF-THEN rules, of which association rules can be regarded as a special case. In [19] rules are clustered according to their coverage of the data, rather than their thematic similarity, using an euclidean metric. In our opinion, however, thematic separation is more adequate for rule set exploration, since the user/analyst will have keep track of the seen and unseen subsets more easily. Data coverage based separation is a good tool for identifying complementary and redundant rules. It is not clear, however, how the result of hierarchical clustering is used.

The combination of association rules and clustering has also been explored before but with different objectives. In [14], association rules with the same consequent and involving exactly the same two attributes in the antecedent are clustered according to a distance function defined in that two-dimensional space. Again, clustering of the rules is dictated by the rules' coverage. A representative rule can be obtained by generalizing the antecedents of the rules in a given cluster. The set of representative rules summarizes the two-dimensional association rules and segment the data. This approach is clearly different from our purposes and, despite its merits, is not appropriate for the interactive exploration

of sets of generic association rules.

A few other works have somehow studied the problem of grouping association rules [24][25], but, to our knowledge, there is no previous proposal for thematic clustering that clearly fits our post-processing aims.

13 Conclusion

Hierarchical clustering of association rules can be used to separate large rule sets thematically. We have defined the requirements for good thematic separation and have shown how these requirements are met by hierarchical clustering, using the binary distance metric between rules and the average linkage strategy for aggregating clusters. The defined requirements also determine the ideal number of clusters for a given set of rules, in terms of clusters separability and integrity.

The division of a rule set into smaller thematic subsets is useful for exploration and post processing. Rule sets of different themes which share no items can be explored independently. This reduces the effort of the data analyst when looking for objectively or subjectively interesting rules. Even if some items are shared between different clusters, the dissimilarity of rules in different clusters is maximized. The division of the rule set can also be done iteratively, if the resulting subsets are considered too large by the analyst or the post processing environment designer.

The rule set can be summarized by choosing a rule, for each cluster, that represents the cluster. The summary set of representative rules allows the analyst to view, in a glance, the thematic divisions of the subsets.

The computed distance between the association rules can be used to measure the lexical distinctness of each rule. This measure, that we have defined according to reasonable requirements, is useful to identify rules that involve items that do not appear frequently in other rules.

In the future, we intend to fully integrate thematic clustering into the PEAR post-processing environment, by providing sets of thematically representative rules as index pages, i.e., starting points of the exploration of the association rule space.

Other possible developments of our work include the study of the robustness of the clustering results with respect to minimal support, as well as the possibility of using clustering techniques that allow overlapping between clusters.

14 Acknowledgements

Special thanks to Paulo Jorge Azevedo from the Departamento de Informática de Universidade do Minho, for making the Caren system available and always adequate to my research needs, and to Carlos Soares from

LIACC-FEP, for his comments on this paper.

References

- [1] J.-M. Adamo, *Data Mining for Association Rules and Sequential Patterns*, Springer-Verlag, 2001.
- [2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, *Fast Discovery of Association Rules*, Advances in Knowledge Discovery and Data Mining: 307-328. 1996.
- [3] P. J. Azevedo, *CAREN A Java based Apriori Implementation for Classification Purposes*, Technical Report, Departamento de Informática, Universidade do Minho, <http://www.di.uminho.pt/~pja/class/caren.html>, 2003.
- [4] R. Bayardo, R. Agrawal, *Mining the Most Interesting Rules*, In Proc. of the 5th International Conference on Knowledge Discovery and Data Mining, KDD - 99, 145-153, 2003
- [5] M. J. A. Berry, and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.
- [6] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, *Dynamic itemset counting and implication rules for market basket data*, SIGMOD Record (ACM Special Interest Group on Management of Data), 26 (2) :255, 1997. <http://citeseer.nj.nec.com/brin97dynamic.html>
- [7] J. Han, Y. Fu, W. Wang, K. Koperski, O. Zaiane, *DMQL: A Data Mining Query Language for Relational Databases*, in Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996.
- [8] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Series in Statistics, Springer, 2001.
- [9] A. Jorge, M. A. Alves, P. J. Azevedo, *Recommendation With Association Rules: A Web Mining Application*, in Proceedings of Data Mining and Warehousing, Conference of Information Society 2002, Eds. D. Mladenic and M. Grobelnik, Josef Stefan Institute, 2002.
- [10] A. Jorge, J. Poças, P. J. Azevedo, *Post-processing operators for browsing large sets of association rules*, in Proceedings of Discovery Science 02, Luebeck, Germany, LNCS 2534, Eds. Steffen Lange, Ken Satoh, Carl H. Smith, Springer-Verlag, 2002.
- [11] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. Verkamo, *Finding interesting rules from large sets of discovered association rules*, in R. Nabil et al., editors, Proceedings of 3rd International Conference on Information and Knowledge Management, pp. 401-407, 1994.
- [12] R. Ihaka and R. Gentleman, *R: A Language for Data Analysis and Graphics*, Journal of Computational Graphics and Statistics, Vol. 5, N. 3, pp. 299-314, 1996.
- [13] B. Liu, W. Hsu, and Y. Ma, *Pruning and Summarizing the Discovered Associations*, In Proc. of the 5th International Conference on Knowledge Discovery and Data Mining, KDD-99, 125-134, 1999.
- [14] B. Lent, A. Swami, J. Widom, *Clustering Association Rules*, in Alex Gray, Per-Åke Larson (Eds.): Proc. of the Thirteenth International Conference on Data Engineering, ICDE 97 Birmingham U.K. IEEE Computer Society, 1997.
- [15] D. D. Lewis, *Reuters-21578 text categorization test collection*, <http://www.daviddlewis.com/resources/test-collections/reuters21578/>, consulted 2003.
- [16] Y. Ma, B. Liu, K. Wong, *Web for Data Mining: Organizing and Interpreting the Discovered Rules* Using the Web, School, SIGKDD Explorations, ACM SIGKDD, Volume 2, Issue 1, July 2000.
- [17] R. Meo, G. Psaila, S. Ceri, *A new SQL-like operator for mining association rules*, in T.M. Vijayaraman et al., editors, Proceedings of the 22nd International Conference on very Large Data Bases, pp. 122-123, 1996.
- [18] R-Project : <http://www.r-project.org>.
- [19] P. Riddle, R. Fresnedo, and D. Newman, *Framework for a generic knowledge discovery toolkit*, In Preliminary papers of the Fifth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale, Florida, January 1995.
- [20] A. Savasere, E. Omiecinski, and S. Navathe, *An efficient algorithm for mining association rules in large databases*, Proc. of 21st Intl. Conf. on Very Large Databases (VLDB), 1995.
- [21] A. Silberschatz, and A. Tuzhilin *On Subjective Measure of Interestingness in Knowledge Discovery*, in Proceedings of the First International Conference on Knowledge Discovery and Data Mining, pp. 275-281, 1995.
- [22] P-N. Tan, V. Kumar, *Interestingness measures for association patterns: a perspective*, in Proceedings of the Workshop on Post-processing in Machine Learning and Data Mining, associated to KDD 2000.
- [23] H. Toivonen, *Sampling large databases for association rules*, Proc. of 22nd Intl. Conf. on Very Large Databases (VLDB), 1996. <http://citeseer.nj.nec.com/toivonen96sampling.html>
- [24] A. Tuzhilin, and G. Adomavicius, *Handling very large numbers of association rules in the analysis of microarray data*, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada: 396-404, 2002
- [25] K. Wang, S.H.W. Tay, B. and Liu, *Interestingness-based interval merger for numeric association rules*, In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), August, 1998.
- [26] C. Zhang, and S. Zhang, *Association Rule Mining: Models and Algorithms*, Lecture Notes in Artificial Intelligence, Vol. 2307, Springer-Verlag, 2002.